**Zhiyu Wu**

Zhiyuwu2@illinois.edu • [**https://xzzwzy.github.io/**](https://xzzwzy.github.io/) • 7345968166 • 410 N Lincoln Ave

**EDUCATION**

**University of Illinois Urbana-Champaign** **Champaign, IL**

*M.S. Computer Science Expected May 2026*

**University of Michigan** **Ann Arbor, MI**

*B.S.E. Computer Engineering (Dual Degree) August 2022 – May 2024*

GPA: 3.58 / 4.00

**Shanghai Jiao Tong University** **Shanghai, China**

*B.S.E. Electrical and Computer Engineering (Dual Degree) Sept. 2020 – August 2024*

GPA: 3.75 / 4.00

Coursework: Computer Architecture, Operating Systems, Computer Network, Machine Learning, Embedded Systems

**PROJECT EXPERIENCE**

**Symbiotic Lab/ML.ENERGY.LEADERBOARD Team Ann Arbor, MI**

*Developer May 2023 – Sept. 2023*

* Developed the ML.ENERGY Leaderboard, an open-source platform for benchmarking the energy efficiency and NLP performance of LLM models.
* Defined performance metrics and implemented scripts for optimized batched inference to ensure accurate measurement.
* Contributed to the online Chatbot Arena which gathers data on models' energy consumption and performance.

**RESEARCH EXPERIENCE**

**Research Assistant in GAEA Lab Champaign, IL**

*Supervisor: Fan Lai July 2024 - present*

* Classify LLM service into latency-sensitive, throughput-intensive, and bulk inference based on system objectives.
* Define Service Level Objectives (SLO) for each type, considering user experience, job completion time (JCT), throughput and “co-flow” completion time, correspondingly.
* Develop an SLO-aware scheduling policy using length prediction to optimize JCT. The policy integrates DAG scheduling and two-dimensional knapsack scheduling for efficient resource allocation, ensuring SLOs are met across request types.

**Research Assistant in Symbiotic Lab Ann Arbor, MI**

*Supervisor: Mosharaf Chowdhury May 2023 – April 2024*

* Identified that in LLM text-streaming services, systems must generate faster than user reading speed to enhance user experience, addressing gaps in previous metrics.
* Defined Quality of Experience (QoE) in LLM serving by tracking each step of text generation and monitoring the overall user experience throughout the entire streaming process.
* Formulated the problem as a knapsack optimization and developed a scheduling algorithm to maximize QoE in online LLM serving.
* Built Andes, an LLM serving system on top of vLLM, integrating the scheduling algorithm to enhance QoE in real-time LLM services.
* Co-authored the paper “Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services” as the second author.

**SKILLS**

***Computer****:* C++, C, Python, Rust, Pytorch, CUDA, System Verilog, Embedded C/Rust, Linux, MATLAB, Git, LaTeX

**HONORS**

Dean List, *Umich* 2023

University Honor, *Umich 2022*

Tang Junyuan Scholarship, *SJTU 2022*

SJTU Undergraduate Excellent Scholarship Class B, *SJTU 2022*